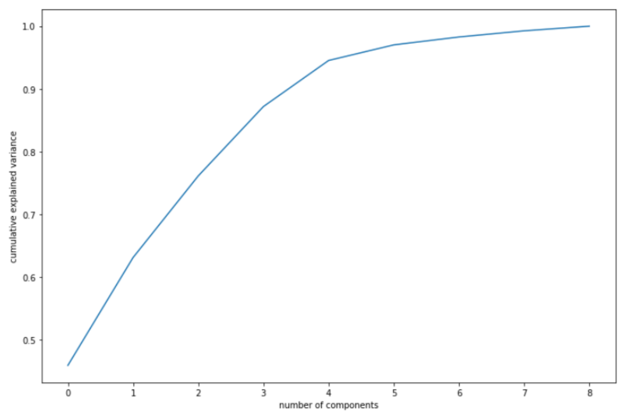
Question 1

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Answer:

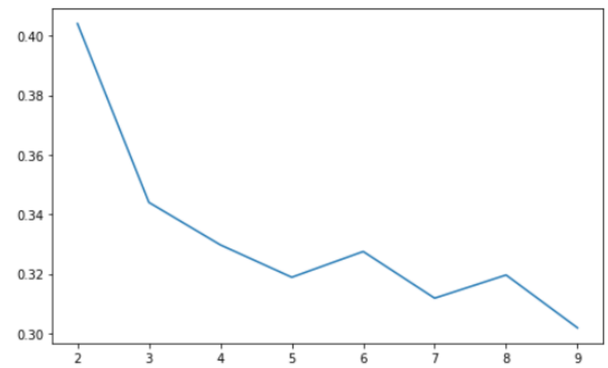
- We first tried to clean the data. We figured out that the data is already cleaned. So we proceeded to do the PCA.
- The PCA needs us to select the optimal number of Principal components. This helps us in the dimensionality reduction in to carry out the clustering process.
- The curve gives an idea of how many PC values one must consider. This is because of the reason that we want to discard the columns/attributes which doesn't participate much in the clustering process.



- Hence, we see that the 80-85% of the values falls under the first 3 attribute values only. So, we choose only 3 Principal components to do the Principal component analysis.
- So, now we have done dimensionality reduction successfully.
- Lets start with the clustering process. The clustering process is of two types. K-means and hierarchical clustering.

- In k-means clustering, we have to pick a number which is K value. This K value is used to define the number of clusters that we want to make with the dataset.
- But, how do we find this K value. ??? we can go ahead and choose this through an analysis.
-

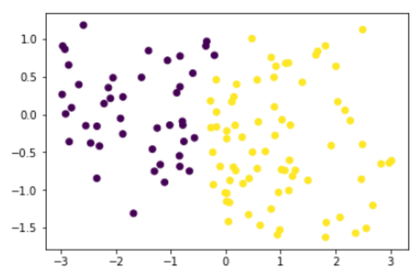
```
In [270]:
fig = plt.figure(figsize = (8,5))
plt.plot(pd.DataFrame(sse_)[0], pd.DataFrame(sse_)[1]);
```



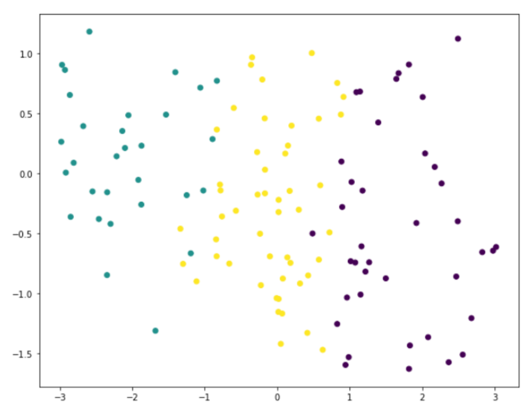
Here, we can see that the graph value is the highest for #2 and then to #3 and then to #5. So these numbers points to the number of ideal clusters one can make from the dataset.

Hence, for k = 2 , the clustering happens this way…

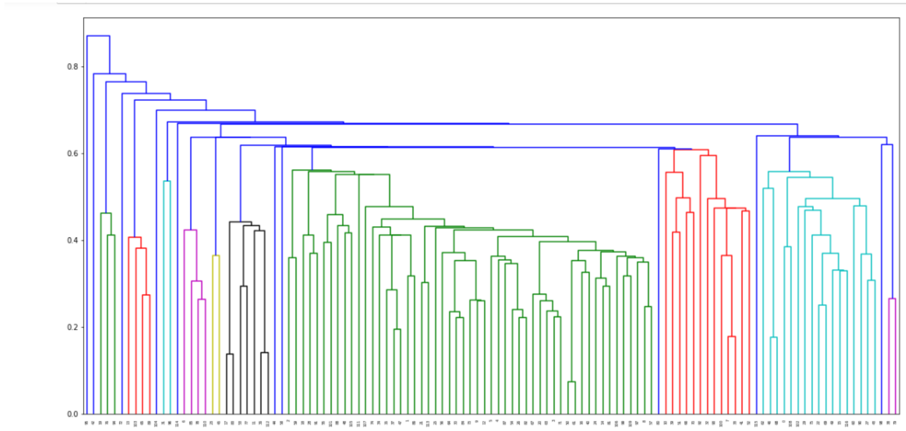
```
In [194]:
plt.scatter(x='PC1',y='PC2',c='ClusterID',data=dat_km)
Out[194]: <matplotlib.collections.PathCollection at 0x1b61f980eb8>
```



Similarly, for k = 3, the data points are classified in the following way.

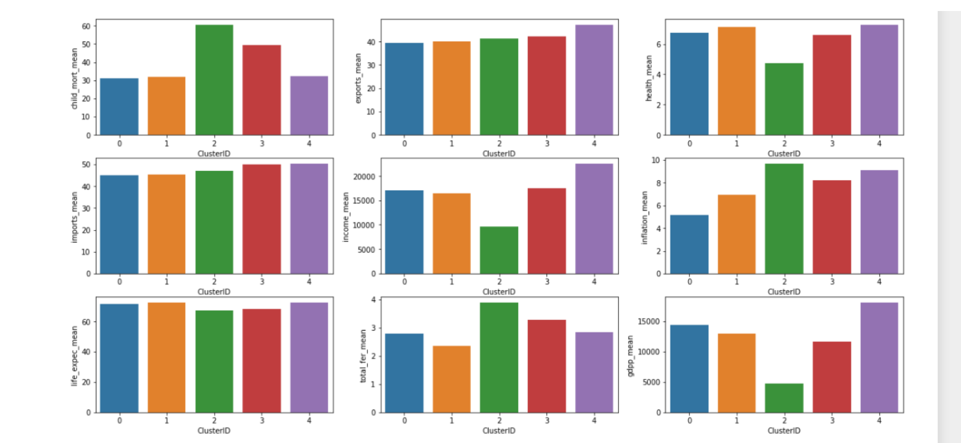


If we'd look at the hierarchical clustering , we can see that the values are linked in the following way..



After analyzing all these clustering, we use the countries and attach with the dataframe that we have been using (al this while, the dataframe consisted of only numerical values in it).

Now, to find out the countries which are the most looked at , we re derived with the clusters which we have visualized in the form of bar plots. (as shown below)



It is to be seen that the countries which belongs to the 5th cluster (i.e. here, it is mentioned to be the 4th cluster in violet) is said to consist of the group of countries which are supposed to be concentrated for funding them by the CEO of HELP foundation.

While sorting the countries by decreasing mortality rate and high GDPP rate, it is come to our conclusion that the countries like :

(i) Haiti

(ii) Sierra Leone

(iii) Chad

(iv) Central African Republic

(v) Mali

Are the top 5 countries which needs to be focused by the CEO of HELP foundation.

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer:

PCA is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to 'find hidden LINEAR correlations' between variables of the dataset. This means that if you have some of the variables in your dataset that are linearly correlated, PCA can find directions that represents your data

- Relies on orthogonal tranformations

Sometimes consider that principal components are orthogonal to the others it's a restriction to find projections with the highest variance

- Large variance = low covariance = high importance

This assumption depends of what problem do you want to solve:
- If you want to compress or remove noise from your dataset this assumption is an advantage
- For mostly any other problem (like Blind Source Separation) it is not useful. Based on Independent Component Analysis theory: uncorrelated is only partly independent.

- mean and covariance doesn't describe some distributions

There are many statistics distributions in which mean and covariance doesn't give relevant information of them. In fact, mean and covariance are used (or could be considered important) for Gaussians.

- scale variant

PCA, as you could've seen, is a rotation trasnformation of your dataset, wich means that doens't affect the scale of your data. It's worth to said also that in PCA you dont normalize your data. That means that if you change the scale of just some of the variables in your data set, you will get different results by applying PCA .

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

K - means clustering	**Hierarchical clustering**
We define the number of clustering in the clustering process (in terms of k value)	We create a dendogram and slaught the dendogram tree into horizontal lines and fetch any number of clusters
With k-Means clustering, you need to have a sense ahead-of-time what your desired number of clusters is (this is the 'k' value). Also, k-means will often give unintuitive results if (a) your data is not well-separated into sphere-like clusters, (b) you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or (c) you have weird initial values for your cluster centroids	hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group.
Simple to construct	Computation expensive
Less time consuming to construct the clusters	Comparatively high time consuming to construct clusters